Privacy and Security Concerns when Social Scientists Work With Administrative and Operational Data

Simson L. Garfinkel

ABSTRACT

Social science is transitioning from working with "designated data" to "organic data." Designated data (also called "made data") are the traditional input of social science—data collected through experiments and surveys. Today the shift is to organic data (also called "found data"), such as administrative data not collected for research purposes, data collected from online social networks, and from large-scale sensor networks. To use organic data in a way that is both fair and accurate, social scientists need to involve computer scientists and, more broadly, information and communications technology (ICT) professionals to obtain, transfer, wrangle, organize, and store massive amounts of organic data so that it can be used as a basis of objective research. Otherwise, researchers risk conclusions that are not widely applicable, and may not even be correct.

The shift to organic data requires significant methodological innovations, and not just because of its size. Because of its diversity, special efforts must be taken to make organic data findable by the broad range of potential users. In some cases, advanced formal privacy techniques such as differential privacy and secure multi-party computation are needed to work with organic data in a manner that is ethically and logistically permissible. Efforts are also required to make studies involving organic data transparent and replicable.

Beyond organic data, our society's growing reliance on ICT is creating opportunities for social scientists to create designated data on scales never possible. Today there are technical infrastructures that make it possible (and common) for a single investigator to engage thousands or even hundreds of thousands of individuals in an experiment. Looking forward, there are opportunities to work with designated data at the scale of 10^6 to 10^8 individuals—for example, by making deliberate changes to the technical infrastructure on which these individuals rely. Experiments have already been conducted showing that privileged social science investigators can covertly manipulate the emotions of people and change the outcome of elections. These experiments will require close partnerships with ICT professionals to assure technical accuracy and scientific validity. Moving forward, social scientists and ICT professionals must develop both appropriate technical controls and ethical frameworks to minimize the risk of these experiments to both the research participants and society at large.

Large-scale information and communications technology (ICT) systems create new opportunities for social scientists; realizing this potential requires strong partnerships between social scientists and computer scientists. In some cases, social scientists can take advantage of off-the-self technology that has been developed over the past decade and is ready to be deployed. In others, social scientists can work with computer scientists to deploy approaches that have been demonstrated in the lab, but still require significant engineering and customization before they can be used more broadly.

Introduction: The Potential of Organic data

For nearly a century, social scientists have worked with data from surveys and statistical agencies. In both cases, these data were typically collected under a promise of confidentiality—sometimes mandated by law and subject to disclosure limitation.[1]

[1] For information on statistical disclosure limitation within the US Government, see Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology,

Today social scientists are supplementing these traditional "designated data" (also called "made data") with "organic data" (also called "found data"), such as administrative records and operational data generated as part of our digitized society.[2] Organic data is plentiful creates exciting new research possibilities, but it may also pose challenges in fairness and accuracy. That's because organic data may contain unknown biases not present in designated data, and may fail to represent various vulnerable populations. And while organic data opens up a potential for important new research projects and evidence-based policy making, the use of such data pose privacy and security challenges as well.[3,4]

One of the premiere examples of organic data used for statistical purposes is the Longitudinal Employer-Household Dynamics (LEHD) Program,[5] operated by the US Census Bureau. LEHD combines traditional census data products and surveys with administrative records about workers and employers to create high-resolution data products that were previously impossible to conceive. The LEHD online tool OnTheMap[6] contains detailed information regarding commuting patterns of workers throughout the 50 states and the District of Columbia, allowing transportation planners to understand the source of traffic and congestion and to make accurate predictions about the impact of new transit options. Another online tool, OnTheMap for Emergency Management,[7] "shows potential impact on jobs/workers and population for hurricanes, tropical storms, fires, floods, snow and freezing rain probability and disaster declaration areas. Real-time geographic data of disaster events are automatically updated."[8]

There's also a growing opportunity to use operational data as well. "Operational data" is the highly granular data that are used to operate systems. For example, whereas OnTheMap considers payroll tax data to infer where people work, commuting patterns could also be

---

Federal Committee on Statistical Methodology, December 2005. https://fcsm.sites.usa.gov/reports/policy-wp/. Current methods of statistical disclosure control are summarized well by George T. Duncan, Mark Elliot, Gonzalez Juan Jose Salazar. *Statistical Confidentiality: Principles and Practice*; Springer Science 2011. For a history of Statistical Disclosure Limitation, see *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, Editors; Panel on Confidentiality and Data Access, National Research Council, ISBN: 0-309-57611-3, 288 pages. http://www.nap.edu/catalog/2122/

[2] "Designated Data" and "Organic Data," *Director's Blog,* Director Robert Groves, US Census Bureau, May 31, 2011. http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/

[3] Lane et al. (eds) Privacy, Big Data and the Public Good: Frameworks for Engagement. Cambridge University Press. 2014

[4] Robert M. Groves and Brian A. Harris-Kojetin, Editors, Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy, National Academies Press, 2017. http://www.nap.edu/24652

[5] LEHD Data, Center for Economic Studies (CES), US Census Bureau. https://www.census.gov/ces/dataproducts/lehddata.html

[6] http://onthemap.ces.census.gov/

[7] https://onthemap.ces.census.gov/em/

[8] LED (Local Employment Dynamics) New Data from the States and the U.S. Census Bureau, 2015. http://lehd.ces.census.gov/doc/LEDonepager_2015.pdf

inferred with real-time trip data collected from electronic toll payment systems (e.g. E-ZPass[9]) or Internet-enabled smart phones.[10,11] Indeed, today operational data from E-ZPass and smart phones are collected and used on a large-scale basis to detect surface traffic flow patterns and display traffic on apps like Google Maps, but not (as far as we know) to create statistical products.

For policy makers and those in the social sciences, there are several key differences between research with designated data and organic data:

- Organic data tends to be larger and more complex than designated data.[12]

- Organic data has the appearance of better coverage, because the data are collected as part of program administration and people cannot opt-out.[13] However, organic data may systematically miss certain populations, or may be edited by unknown processes before it reaches researchers. For this reason, the data accuracy may be illusionary.

- Because organic data might be collected covertly as part of a person's ordinary activities, such as performing internet searches or driving around the city, handling and using organic data may require more attention to privacy issues than using designated data.

- Because of these privacy issues, there may be a greater need to remove identifiers in organic data. The process of removing identifiers and other identifying information is called "de-identification." De-identified data is not necessarily safe to publicly release, however, as de-identified data can be re-identified.

- Whereas designated data are created for statistical analysis, organic data are created for other purposes. Unlike designated data, "researchers generally have no input into the design, structure and content of administrative social science data."[14] As a result, organic

---

[9] Kashmir Hill, "E-ZPasses Get Read All Over New York (Not Just At Toll Booths)," Forbes, Sept. 12, 2013. http://www.forbes.com/sites/kashmirhill/2013/09/12/e-zpasses-get-read-all-over-new-york-not-just-at-toll-booths/

[10] Jungkeun Yoon, Brian Noble, and Mingyan Liu. 2007. Surface street traffic estimation. In *Proceedings of the 5th international conference on Mobile systems, applications and services* (MobiSys '07). ACM, New York, NY, USA, 220-232. DOI=http://dx.doi.org.proxy.library.georgetown.edu/10.1145/1247660.1247686

[11] Mingqi Lv, Ling Chen, Gencai Chen, and Daqiang Zhang. 2014. Detecting traffic congestions using cell phone accelerometers. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (UbiComp '14 Adjunct). ACM, New York, NY, USA, 107-110. DOI=http://dx.doi.org.proxy.library.georgetown.edu/10.1145/2638728.2638744

[12] L. Taylor, R. Schroeder, E. Meyer, Emerging practices and perspectives on Big Data analysis in economics: bigger and better or more of the same? Big Data Soc., 1 (2014) 2053951714536877

[13] L. Einav, J.D. Levin, The Data Revolution and Economic Analysis, National Bureau of Economic Research (2013)

[14] Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben, "The role of administrative data in the big data revolution in social science research," *Social Science Research* 59 (2016), Special Issue on Big Data in the Social Sciences, pp. 1–12, Elsevier. http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

data needs to be carefully evaluated to see if it is suitable for the intended research and statistical purposes.

- Because it wasn't created for research, organic data may have errors and uncertainty that might be known to the primary users of the data, but are frequently not known to the secondary researchers. These "unknown unknowns"[15,16] can be different from traditional sources of error and uncertainty.

- Because it is collected at large scale and for purposes other than scientific research, administrative and operational data can take researchers to places where they don't belong, giving them access to secrets that are outside the scope of legitimate scientific inquiry, crossing into the realm of the private or even the prurient. Administrative controls such as two-person review of queries and unforgeable audit logs can act as a deterrent against abuse by researchers.

Tools like OnTheMap show that important policy questions can answered with carefully designed tools that use a combination of designated data (surveys) and organic data (tax records). OnTheMap demonstrates that these data products can be generated in a way that preserve confidentiality. (In the case of OnTheMap, personal privacy is protected through the infusion of dynamically consistent noise[17] so that the individual contributions of a specific person or company to data cannot be discerned.)

OnTheMap also shows that efforts to use organic data by social scientists need to address several key issues:

1. Administrative and operational datasets will frequently contain some kind of identifiers. As a result, there frequently needs to be some way to de-identify the data before it is released to researchers.

2. Data confidentiality must be protected, even as the data are made more widely available.

3. Data provenance must be tracked.

4. Researchers should consider advanced techniques for privacy-preserving data processing, such as secure multi-party computation.

5. Because it is a short walk from collecting data on a large number of people to collecting data after minor interventions, social science researchers should acquaint themselves

---

[15] Thomas P. Coakley, *Command and Control for War and Peace*, National Defense University Press, Washington DC, 1992, Chapter 5, "An Impossible Dream: Information That is Complete, True, and Up-to-date," p. 136.
[16] Oettinger, Anthony G., "Future Innovations: The Endless Adventure," *Bulletin of the Association for Information Science and Technology*, 27:2, December/January 2001, pp. 10-15. http://onlinelibrary.wiley.com/doi/10.1002/bult.190/full
[17] Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. Red Hook, N.Y.: Curran Associates.

with existing ethical frameworks for working with human subjects and emerging norms for experimentation that involves people using computing systems.

The Challenge of De-identification

Ethical and logistical complications can arise when working with organic data. For example, there may be more privacy concerns, because people tend to edit themselves when filling out a survey—facts that might be included in the data dump of an operational system. Unlike designated data, data subjects generally haven't been given *notice* that their data will be used for research, and they haven't given their *consent* to be involved in research studies.

The lack of informed consent poses a problem for researchers in the United States operating with federal funding. Such researchers are covered under the Common Rule,[18] which requires that research involving human subjects (or identifiable private data about humans) be approved by an accredited institutional review board (IRB), and that they give informed consent. Although IRBs can waive the informed consent provisions,[19] many researchers find it easier to *de-identify* organic data—that is, to remove the identifiers, so that the Common Rule does not apply.

Today de-identification is widely practiced. For example, the US Department of Health and Human Service (HHS) Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[20] states that de-identified health information is not considered protected health information (PHI) and, as such, is not legally protected.[21] As a result, de-identified health information is widely used for both medical research and marketing. Likewise, the Family Educational Rights and Privacy Act (FERPA) specifically allows de-identified data to be "shared without the consent required by FERPA (34 CFR §99.30) with any party for any purpose, including parents, general public, and researchers (34 CFR §99.31(b)(1))."[22]

Most de-identification protocols require classifying the variables in a dataset as *direct identifiers* (such as names and phone numbers) that can directly identify a data subject; *indirect identifiers* or *quasi-identifiers* (such as a person's age or height) that can be used to narrow down a data subject from a set of possible subjects; and *non-identifying sensitive values* (such as a person's diagnosis). Once the variables are classified, direct identifiers are removed and quasi-identifiers

[18] 45 CFR 46
[19] See 45 CFR 46.116(d) for the criteria for waiving informed consent. Note that the Food and Drug Administration's (FDA's) version of the Common Rule (21 CFR 50 and 56) and the Department of Health and Human Services (DHSS) regulations (21 CFR 46) are more restrictive regarding such waivers.
[20] 45 CFR 160 and 45 CFR 164 subparts A and E; see also https://www.hhs.gov/hipaa/for-professionals/privacy/.
[21] See also "How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?" US Department of Health and Human Services National Institutes of Health, https://privacyruleandresearch.nih.gov/pr_08.asp. Last updated February 2, 2007.
[22] "Data De-Identification: An Overview of Basic Terms," Privacy Technology Assistance Center, US Department of Education http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf, last updated May 2013.

are manipulated so that individuals can no longer be reliably identified. This kind of de-identification protocol relies on assumptions about the background information available to a person attempting to an individual who might attempt to re-identify records in a dataset.

In recent years, de-identification techniques have come under attack by computer science researchers, who have shown that many datasets that had been de-identified can be re-identified—that is, individual data records can be matched back up with the true identities.[23]

As an alternative to releasing de-identified data, social scientists can create interactive query systems that allow limited access to data without risk of compromising the privacy of individuals within the dataset.[24] Alternatively, data can be used to create synthetic datasets that preserve some relationships between variables while preventing the re-identification of specific individuals.

As social science researchers transition from working with "designated data" to "organic data," they need to partner with computer scientists to deploy new approaches for working with data in ways that provide protection appropriate to the sensitivity of these datasets. Fortunately, computer science is up to the task, with a wide range of techniques that have been developed over the past decade that can protect the confidentiality of data while unlocking its potential for use by researchers.

Key Points:

- De-identification protocols typically remove direct-identifiers and may manipulate quasi-identifiers.
- De-identified data can be re-identified.
- The risk of re-identification can be minimized through the use of formal privacy methods, such as interactive query systems for data access, or through the creation of synthetic datasets.

Keeping Data Confidential

In 1986, John Diebold, one of the pioneers of the computer age explained how transactional data containing place and time information could become far more sensitive than first apparent. The case involved a bank that, in 1979, "had recently installed an automatic teller machine network and noticed 'that an unusual number of withdrawals were being made every night between midnight and 2:00 a.m," Diebold wrote. "Suspecting foul play, the bank hired detectives to look into the matter. It turns out that many of the late-night customers were

[23] For a discussion of several high-profile attacks on de-identified data sets, see Simson L. Garfinkel, NISTIR 8053: De-Identification of Personal Information, National Institute of Standards and Technology Interagency Report 8053, October 2015. http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf
[24] Frank McSherry, Privacy Integrated Queries, Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, https://www.microsoft.com/en-us/research/publication/privacy-integrated-queries/.

withdrawing cash on their way to a local red light district!"[25] An article about the incident that appeared in the Knight News Service observed: "there's a bank someplace in America that knows which of its customers paid a hooker last night."[26]

Diebold's story vividly demonstrated that administrative and operational data can pose challenges resulting from both sensitivity and scale when compared with traditional social science data. With traditional surveys and laboratory experiments, respondents who don't want to take part in a research project have many means for opting out—for example, by not filling out a survey, or by not answering specific questions. But when it comes to administrative and operational data, opting out is harder. It's unlikely that a person will opt out of using an automated teller in 1979 or a cell phone today to avoid participating in some economist's research project. This inability of opting out is one of the things that make these data sets so particularly attractive: they hold the appearance of both accuracy and completeness, something missing from traditional surveys. But the inability of opting out also means that researchers who have access to such data need to adopt strong controls for preserving data confidentiality, least the researchers provoke a public backlash. (Researchers also need to realize that this appearance of accuracy and completeness is frequently an illusion, just as it was unlikely that every person withdrawing cash from that ATM in 1979 between midnight and 2:00am was actually visiting a prostitute.)

Countless news reports over the past two decades has shown that operational and administrative data is a powerful magnet for computer hackers, criminals, and even espionage. As such, researchers have an obligation to protect the data using mechanisms that are appropriate to their sensitivity. Researchers who have been trained and become accustomed to working with publicly available datasets may require significant retraining and retooling before working with datasets containing sensitive information.

At a minimum, sensitive data should be stored on devices featuring full disk encryption to minimize the chance of data compromise when equipment is decommissioned[27] and to protect against the risk of compromise in the event of equipment theft. There should also be written, audited procedures to control devices that store data, including computers, hard drives, and portable storage media, to assure that these devices are appropriately sanitized or physically destroyed when they are retired from service.

But encryption alone is insufficient to assure the confidentiality of sensitive data. Sensitive data may become the targets of organized hacking efforts by criminals, hacktivists and even foreign governments. Motives for attack may include financial gain, the desire to embarrass the

[25] John Diebold, in James Finn and Leonard R. Sussman, eds. *Today's American: How Free?* New York: Freedom House, 1986. p. 111
[26] ibid.
[27] S. L. Garfinkel and A. Shelat, "Remembrance of data passed: a study of disk sanitization practices," in *IEEE Security & Privacy*, vol. 1, no. 1, pp. 17-27, Jan.-Feb. 2003.

organization that provided the data, and obtaining information for attacking the data subjects themselves. [28]

To protect against these threats, researchers should use computers running up-to-date operating systems, with up-to-date anti-virus/anti-malware systems installed, and researchers should have mandatory and ongoing cybersecurity training. Systems should be professionally managed, and should have network connections that are monitored by a competent entity, such as a managed security provider. Systems holding confidential data should only be accessible with two-factor authentication.

Some data may be so sensitive that they should only reside on systems that are physically disconnected from the Internet—for example, a stand-alone desktop or an "air-gapped" network. An air-gapped network might consist of a storage server and five or ten workstations connected to a gigabit switch. To copy software or data from the Internet to the private network, it is burned onto a DVD. The DVD is then taken to a stand-alone computer where it is virus-scanned. If the scan is successful, the DVD is then taken to the air-gapped network and copied to the server. Removing information follows the same procedure in reverse. Clearly, building and maintaining an air-gapped network is complicated—and easy to get wrong.[29] Sadly, there are few resources available with step-by-step instructions on how to create and maintain such networks.[30]

A workable middle ground between networks that are full-connected to the Internet and air-gapped networks is to build a so-called *secure enclave*, in which the data resides on secure servers that can only be reached through audited, controlled, mechanisms. A *physical secure enclave* might be a physical facility with locks, a guard, and computers that can reach the research data but not connect to the Internet. This is the approach taken by the Federal Statiatical Research Data Centers and the NORC Data Enclave. A *virtual secure enclave* might be a computer system that can only be reached by secure low-bandwidth interconnections. In either case, users can enter the secure enclave and run queries and see the results, but they can't export the data until it has undergone some kind of review process, ideally one that is formally vetted by a disclosure review board. Untrusted users may even be prohibited from seeing the results of their queries until a formal review is completed.

---

[28] See Federal Court: Canada (2007) Mike Gordon v. the Minister of Health and the Privacy Commissioner of Canada: Memorandum of Fact and Law of the Privacy Commissioner of Canada. Federal Court, as referenced in Khaled El Emam, Elizabeth Jonker, Luk Arbuckle and Bradley Malin, A Systematic Review of Re-Identification Attacks on Health Data, PLoS ONE, 6:12, December 2011. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028071

[29] Eric Byres. 2013. The air gap: SCADA's enduring security myth. Commun. ACM 56, 8 (August 2013), 29-31. DOI=http://dx.doi.org/10.1145/2492007.2492018

[30] See, for example, Bruce Schneier, "Air Gaps," Schneier on Security, https://www.schneier.com/blog/archives/2013/10/air_gaps.html. One of the main challenges in operating an air-gapped network is installing updates to software packages, see "Setting up OS Deployment in an air-gapped network," IBM Knowledge Center, http://www.ibm.com/support/knowledgecenter/SS63NW_9.2.0/com.ibm.tem.life.doc_9.2/Lifecycle_Man/OSD_Users_Guide/c_osd_setup_airgap.html

Security techniques are also needed to protect data from the researchers themselves. Strong internal security measures can protect data subjects from inappropriate accesses to their data, and serve as a deterrent to help keep researchers on the correct side of proper ethical conduct. For example, column-level encryption[31] locks some fields of the database to some users, preventing unrestricted access.

Properly implemented, encryption blocks inadvertent or unauthorized access to sensitive information, while not eliminating the possibility of having authorized individuals access the encrypted data using appropriate mechanisms. In general, access control technologies such as encryption force designers to think out and plan for what kind of access is needed in advance. Pilot studies, end-to-end tests and "dress rehearsals" can help identify unanticipated usability challenges that might impact data quality, such as the use of encrypted fields that makes it difficult to evaluate the quality of a database linkage.

Because of its richness and broad coverage, operational and administrative data can tempt individuals in ways that data collected for traditional statistical projects do not. For example, in 2013, *The Wall Street Journal* featured an article detailing how analysts at the National Security Agency used intelligence collection systems to spy on their love interests.[32] Similar abuses of official information systems occur at the state and local level: in 2012 Anne Marie Rasmusson, a former police officer, was awarded $1M in her federal invasion-of-privacy lawsuit against the cities of Minneapolis and St. Paul, after 104 police offers illegally accessed her photo and other driver's license data.[33]

Secure audit logs to detect inappropriate data searches and browsing.[34] But care must be taken so that reasonable security measures do not inadvertently deter quality research. As Diebold's story demonstrates, sometimes the very artifacts that make data interesting that also make data sensitive. Minor inconsistencies and variations can attract the attention of an inquisitive mind; drilling down reveals secrets: there needs to be a way for researchers to explore data, conduct

---

[31] For example, T. Ge and S. Zdonik, "Fast, Secure Encryption for Indexing in a Column-Oriented DBMS," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, 2007, pp. 676-685. A general overview for less-technical readers can be found in Ashvin Kamaraju, "Database encryption demystified: Four common misconceptions," ZDNet February 9, 2012. http://www.zdnet.com/article/database-encryption-demystified-four-common-misconceptions/

[32] Siobhan Gorman, "NSA Officers Spy on Love Interests," *The Wall Street Journal,* Aug. 23, 2013. http://blogs.wsj.com/washwire/2013/08/23/nsa-officers-sometimes-spy-on-love-interests/. See also Andrea Peterson, "LOVEINT: When NSA officers use their spying power on love interests," *The Washington Post*, August 24, 2013. https://www.washingtonpost.com/news/the-switch/wp/2013/08/24/loveint-when-nsa-officers-use-their-spying-power-on-love-interests/?utm_term=.7a02a2ebdf1d

[33] Kim Zetter, "Female Cop Gets $1 Million After Colleagues Trolled Database to Peek at Her Pic," Wired, November 5, 2012. https://www.wired.com/2012/11/payout-for-cop-database-abuse/. See also Jessica Lussenhop, "Is Anne Marie Rasmusson too hot to have a driver's license?", City Pages, February 22, 2012. http://www.citypages.com/news/is-anne-marie-rasmusson-too-hot-to-have-a-drivers-license-6755567

[34] For example, see Gunnar Hartung. 2016. Secure Audit Logs with Verifiable Excerpts. In *Proceedings of the RSA Conference on Topics in Cryptology - CT-RSA 2016 - Volume 9610*, Kazue Sako (Ed.), Vol. 9610. Springer-Verlag New York, Inc., New York, NY, USA, 183-199. DOI: http://dx.doi.org/10.1007/978-3-319-29485-8_11; Di Ma and Gene Tsudik. 2009. A new approach to secure logging. *Trans. Storage* 5, 1, Article 2 (March 2009), 21 pages. DOI=http://dx.doi.org/10.1145/1502777.1502779.

spot checks, and look for potentially interesting artifacts, without setting off alarms and trashing careers.

Key Points:

- Administrative and operational data can contain information that is sensitive; unlike traditional social science data collected by surveys, there is no obvious way for members of the public to "opt-out." As a result, there is a higher moral obligation on researchers to keep such data confidential.

- Encryption can protect data from unauthorized access by individuals or organizations.

- Encryption can be applied to a device (such as a laptop), a file (such as a database), or a specific column of a database.

- Data can also be compromised by malware and hostile outsiders. One way to prevent compromise is to confine data to computers that are never connected to the Internet, such as stand-alone computers or computers connected to an "air-gapped" network.

- Another way to protect data confidentiality is by placing a "data enclave" that has a limited connection to the Internet. There are many kinds of data enclaves. One kind has a limited connection that allows information such as queries and query results to move back-and-forth, but the connection does not allow entire datasets to be transferred.

- Sensitive data is subject to abuse. Administrative controls such as two-person checks and unforgeable query logs can act as a deterrent against insider abuse.

Data Curation and Data Provenance

To use organic data in ways that are both fair and accurate requires more than good security: it requires a principled approach to data management and data processing.

Administrative and operational data can be messy. Data from production systems frequently contains format inconsistencies—errors that look like typos, for example. Depending on the data source, the inconsistencies may be actual typos, transmission errors, format changes over time, or even real data. Failure to handle such errors can result in systematic bias being introduced into the dataset, which may prime the unwary experimenter for false discovery.

The term *data curation* has emerged to describe the process of managing data through the data lifecycle. Data curation is widely seen as an important task in any organization that relies on data for operations, research, or policy making.

For example, avoiding, detecting and correcting errors requires careful attention to data specifications, formats, and changes in software versions. Thus, capturing information about processing software and runtime environments is an important aspect of data curation. (There are cases of commercial software behaving differently on different operating systems—and even behaving differently if the software is run in the summer, when daylight savings time is in effect, or in the winter, when it is not.)

An added complication is that the data curator might be mistaken, and edit data that is thought to be erroneous, but which, in fact, is not. For these reasons, it is important to record both old and new values when editing.

By its very nature, operational data has the tendency to grow without limit. Researchers must decide what data to keep and what to discard. Uncomfortable when placed in the position of having to decide about data retention, some researchers punt and keep it all. Thus, data tends to increase over time. And because researchers need multiple copies of data to protect against operator error[35] and silent data corruption,[36] storage requirements tend to grow geometrically.

Data Provenance is a term used to describe information about how data are collected, processed, stored and distributed. Provenance can (and should) also document the individuals, systems, and software that collects, processes, stores and distributes data. Provenance thus includes what is commonly thought of as "metadata"—but it also includes information that is not typically captured, such as the name of the person who performed the data analysis, the specific version of the software that was used to perform the analysis, and the steps that were taken to clean the data. Systematically capturing and maintaining provenance is an important part of data curation.

Researchers purchasing or developing large scale storage systems should investigate approaches for automatically capturing provenance and storing it as part of their metadata. Collecting and analyzing provenance can have unexpected benefits, such as helping to understand and improve database performance.[37]

Users of commercial statistics packages tend to underestimate the amount of provenance that is required to reconstruct a finding. For example, it may be necessary to capture both the user's scripts (e.g. a Stata "do-file"), the version of the statistics package, the version of the host operating system, the model number of the computer's microprocessor (CPU), the time and date

---

[35] Aaron B. Brown and David A. Patterson, "To Err is Human," Proceedings of the First Workshop on Evaluating and Architecting System dependabilitY (EASY '01), http://roc.cs.berkeley.edu/papers/easy01.pdf

[36] For information on silent data corruption, see Sumit Narayan, John A. Chandy, Samuel Lang, Philip Carns, and Robert Ross. 2009. Uncovering errors: the cost of detecting silent data corruption. In *Proceedings of the 4th Annual Workshop on Petascale Data Storage* (PDSW '09). ACM, New York, NY, USA, 37-41. DOI=http://dx.doi.org/10.1145/1713072.1713083; David Fiala, Frank Mueller, Christian Engelmann, Rolf Riesen, Kurt Ferreira, and Ron Brightwell. 2012. Detection and correction of silent data corruption for large-scale high-performance computing. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (SC '12). IEEE Computer Society Press, Los Alamitos, CA, USA, Article 78, 12 pages; and Leonardo Bautista Gomez and Franck Cappello. 2014. Detecting silent data corruption through data dynamic monitoring for scientific applications. In *Proceedings of the 19th ACM SIGPLAN symposium on Principles and practice of parallel programming* (PPoPP '14). ACM, New York, NY, USA, 381-382. DOI=http://dx.doi.org/10.1145/2555243.2555279;

[37] Peter Macko, Daniel Margo, and Margo Seltzer. 2013. Performance introspection of graph databases. In *Proceedings of the 6th International Systems and Storage Conference* (SYSTOR '13). ACM, New York, NY, USA, , Article 18 , 10 pages. DOI=10.1145/2485732.2485750 http://doi.acm.org/10.1145/2485732.2485750

that the software was run, the time zone, the amount of random access memory (RAM) installed in the computer, and other seemingly benign information. Having all this information captured, stored and permanently recorded can be vital years later, when future researchers are trying understand why today's results can't be replicated. Even in the short-term, this information can be vital to other scientists seeking to reproduce findings.

It is critical to automatically collect and index provenance to make data findable in a large-scale data research environment. With multiple researchers at multiple institutions, provenance can feed a search engine and respond to queries such as *find data collected between 2010 and 2014 that was processed with R to produce a table of household income vs. immunization rates*. Of course, to be findable, the search interface must be accessible to would-be downstream users. One way to assure this is through a technique called *federated search,*[38] in which multiple search engines are tied together so that a single search request can answered by dozens or hundreds of independent but federated search engines. Federated search can result in results that more both more relevant and more diverse[39] than queries to a single search engine, and approaches to address privacy issues have been developed.[40]

Key Points:

- Provenance documents what happens to data, including collection, processing, dissemination and storage.
- Provenance documents how data are transformed, including the names of the people who perform the transformations, the software that they use, and the process.
- Provenance can be stored with data or separately.
- Provenance can be collected and stored automatically, or manually.
- Carefully collected and indexed, provenance can make data more findable and usable.

Privacy Preserving Data Collection, Processing and Publishing

So far, this chapter has been concerned with techniques developed by computer scientists for securely storing sensitive information and making the results of data processing findable. But computer scientists have also developed techniques over the past three decades that can be used

---

[38] Dong Nguyen, Thomas Demeester, Dolf Trieschnigg, and Djoerd Hiemstra. 2012. Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (CIKM '12). ACM, New York, NY, USA, 1874-1878. DOI=http://dx.doi.org/10.1145/2396761.2398535

[39] Dzung Hong and Luo Si. 2013. Search result diversification in resource selection for federated search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '13). ACM, New York, NY, USA, 613-622. DOI=http://dx.doi.org/10.1145/2484028.2484091

[40] Wei Jiang, Luo Si, and Jing Li. 2007. Protecting source privacy in federated search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '07). ACM, New York, NY, USA, 761-762. DOI=http://dx.doi.org/10.1145/1277741.1277896

A Multiparty Computation Example

Although the solution for two-party protocols is too complicated to present here, there is a simple three-party protocol for a similar problem that is presented below.

Consider the problem *Average Salary Problem,* in which a group of people wish to compute their average salary without revealing their specific salary to each other or to a third party. In this case, we will consider Alice, Bob and Carol, whose salaries are A, B and C, respectively. The three wish to compute value of (A+B+C)/3 without revealing A, B or C to each other.

The solution to this problem is straightforward. Alice, Bob and Carol each chose a random number—R_A, R_B and R_C, respectively. Alice sends to Bob the number A-R_A and to Carol the number R_A. Neither Bob nor Carol have enough information to reconstruct Alice's number A. Bob, meanwhile, sends to Alice the number B-R_B and Carol the number R_B. Carol sends the number C-R_C to Alice and R_C to Bob. Next, each of the players add together the numbers that they have received from the other two: Alice adds the numbers (B-R_B) and (C-R_C) to get (B+C-R_B-R_C), Bob adds the numbers (A-R_A) and R_C to get (A-R_A+R_C), and Carol adds R_A and R_B to get (R_A+R_B). Alice, Bob and Carol now each write their sum on a whiteboard at the front of the room:

Alice writes the single number AA that is (B+C-R_B-R_C).

Bob writes the single number BB that is (A-R_A+R_C).

Carol writes the single number CC that is (R_A+ R_B).

The three numbers are now added and divided by three. This number, (AA+BB+CC)/3 is equal to ((B+C-R_B-R_C) + (A-R_A+R_C) + (R_A+R_B)) / 3 = (A + B + C) / 3 , which is the value that was to be computed! Thus, Alice, Bob and Carol have computed their average salary without revealing their individual salaries to each other. ♦

for collecting, processing, and publishing sensitive information in a manner that preserves privacy.

Many of these techniques find their intellectual heritage in the work of Andrew Yao, a cryptographer who in 1982 introduced the concept of secure two-party computation, also called secure function evaluation. Yao developed a solution to the *Millionaires' Problem*, in which two millionaires, Alice and Bob, engage in a two-person mathematical protocol that lets the them determine who is the richer of the two without reveal their wealth to each other or to a trusted third party.[41] For an example of how such a multi-party computation might take place, please see the Multiparty Computation Example in the box.

[41] Yao, Andrew C. (November 1982). "Protocols for secure computations". *FOCS*. 23rd Annual Symposium on Foundations of Computer Science (FOCS 1982): 160–164. doi:10.1109/SFCS.1982.88.

Since Yao's discovery, many protocols and procedures for various kinds of security and privacy preserving computations have been discovered.

Private information retrieval[42] describes a family of protocols that allow for a user to retrieve data from a server without revealing which item is retrieved. The first scheme was introduced in 1997 by Kushilevitz and Ostrovsky; recently Yi, Paulet and Bertino published a survey of many techniques.[43]

Search on Encrypted Data are techniques that allow for a client to conduct searches on an encrypted database, without revealing the contents of the encrypted documents or the search terms. Early work was done by Song, Wagner and Perrig;[44] follow-up work by others has discovered approaches for searches that are tolerant of minor misspellings,[45] rank keywords;[46] and even find medical imagery.[47]

Oblivious RAM (ORAM), in which information is stored and retrieved from a remote server, but the server is unable to determine what is stored, what is retrieved, or the update patterns.[48]

Cryptographic Voting Protocols,[49] in which participants vote and votes are tallied, but for which it is impossible to determine the vote of any specific voter. Many voting protocols have additional properties, such as the ability of voters to verify that their votes were counted, or the ability to mathematically prove that the votes were properly counted.

---

[42] Kushilevitz, Eyal; Ostrovsky, Rafail (1997). "Replication is not needed: single database, computationally-private information retrieval". *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*. Miami Beach, Florida, USA: IEEE Computer Society. pp. 364–373. ISBN 0-8186-8197-7.

[43] Xun Yi; Russell Paulet; Elisa Bertino, "Private Information Retrieval," in *Private Information Retrieval* , 1, Morgan & Claypool, 2013, pp.114; doi: 10.2200/S00524ED1V01Y201307SPT005

[44] Dawn Xiaoding Song, D. Wagner and A. Perrig, "Practical techniques for searches on encrypted data," *Proceeding 2000 IEEE Symposium on Security and Privacy. S&P 2000*, Berkeley, CA, 2000, pp. 44-55.

[45] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren and W. Lou, "Fuzzy Keyword Search over Encrypted Data in Cloud Computing," *2010 Proceedings IEEE INFOCOM*, San Diego, CA, 2010, pp. 1-5.

[46] C. Wang, N. Cao, J. Li, K. Ren and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," *2010 IEEE 30th International Conference on Distributed Computing Systems*, Genoa, Italy, 2010, pp. 253-262.

[47] J. Yuan, S. Yu and L. Guo, "SEISA: Secure and efficient encrypted image search with access control," *2015 IEEE Conference on Computer Communications (INFOCOM)*, Kowloon, 2015, pp. 2083-2091.

[48] Oded Goldreich. 1987. Towards a theory of software protection and simulation by oblivious RAMs. In Proceedings of the nineteenth annual ACM symposium on Theory of computing (STOC '87), Alfred V. Aho (Ed.). ACM, New York, NY, USA, 182-194.

[49] Chris Karlof, Naveen Sastry, David Wagner, "Cryptographic Voting Protocols: A Systems Perspective," 14th USENIX Security Symposium, 2005.
https://www.usenix.org/legacy/event/sec05/tech/full_papers/karlof/karlof.pdf

Differentially Private Algorithms[50,51], in which queries executed over a dataset, are reported in a way that any specific individual's contribution cannot be inferred with a degree of certainty. Many differentially private algorithms are based on the addition of Laplace noise to the results of queries: by carefully controlling the amount of noise added, these algorithms can produce results that are both reasonable accurate and reasonably privacy preserving.

Collectively, these approaches can be called *formal privacy techniques*, because the privacy assumptions and guarantees are formally stated and privacy loss or protection is mathematically provable. When using a formal privacy technique, is it important to understand how words like *privacy* and *security* are formally defined, since the defined meanings may be subtly different than the colloquial ones.

Formal privacy techniques can be combined with the other privacy protecting techniques discussed earlier in this chapter. For example, an organization that has sensitive data within a data enclave could use an algorithm like Ullman's Private Multiplicative Weights[52] to produce a differentially private synthetic dataset that is publicly distributed. Researchers could use this dataset to develop queries and to perform their initial data analysis. Once the code is working, the researchers could then provide their code to operators of the secure data enclave, who would then run the code on the actual data and review the results for inappropriate disclosures prior to making the results available to the researchers.

Key Points:

- Cryptographic techniques exist that can compute functions over private data such as sums and averages without releasing the actual data to anyone, including a trusted third party.

- A rich tool chest of algorithms exists, but they have not seen wide adoption as of yet.

Found experimental opportunities

Just as organic data creates the possibility for social scientists to perform statistical analyses at a scale never before possible, *found experimental opportunities* makes it possible for social scientists to intervene and conduct experiments on people and societies at a new, grand scale as well—experiments involving thousands, millions or even billions of people. These capabilities have the potential to cause significant disruption to the way that many people do science.

---

[50] Calibrating Noise to Sensitivity in Private Data Analysis by Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith In Theory of Cryptography Conference (TCC), Springer, 2006. DOI=10.1007/11681878_14

[51] The Algorithmic Foundations of Differential Privacy by Cynthia Dwork and Aaron Roth. Foundations and Trends in Theoretical Computer Science. Vol. 9, no. 3–4, pp. 211-407, Aug. 2014. DOI=10.1561/0400000042

[52] Jonathan Ullman. 2015. Private Multiplicative Weights Beyond Linear Queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (PODS '15). ACM, New York, NY, USA, 303-312. DOI=http://dx.doi.org/10.1145/2745754.2745755

Only a few short years ago, recruiting 500 or 1000 people for a survey or study could be a painstaking task. Today there are numerous services such as Mechanical Turk,[53] CrowdFlower[54] and Prolific Academic[55] that let a solitary researcher create a web-based experiment, ship it out to "workers," and get a response within hours. Prices for tasks can be so low—perhaps 20 cents per subject—that a student or even an enthusiastic amateur can fund an experiment on hundreds or thousands of people without having to rely on institutional funds.

Experimenting with ones' own funds can place a faculty researcher in an ethically ambiguous area. The primary mechanism by which human subjects research is regulated in the United States, the Institutional Review Board (IRB) system as defined by the National Research Act[56] and the Common Rule,[57] only apply to research that is federally funded. Although many universities require that all research involving human subjects performed under the auspices of a university be approved by the university's IRB, there is no general requirement that experiments on human subjects, even medical experiments, be approved by a disinterested, objective body. Some academics might think that using their own funds, on their own time, with their own computers, might somehow absolve themselves of the need to get IRB approval.

Individuals who carry out social science research without oversight have the potential to "poison the well" of good will—and even societal tolerance—for social science research in general, just as so-called *push polls,*[58] also known as advocacy polls, have damaged the credibility, effectiveness, and even the legitimacy of traditional public opinion polls. But whereas sending out a push poll to several thousands of homes might costs thousands of dollars, sending out a survey by Mechanical Turk to thousands of respondents might cost less than $100.

Another way to experiment on thousands of people is to package the experiment into an app and upload it to Apple's App Store or Google Play. An ethical experimenter might clearly disclose its purpose, protocol, and obtain the user's permission to proceed—a kind of informed consent. Alternatively, the app might simply offer some sort of compelling feature to its users, and perform the experiment covertly.

Yet another way to experiment on people—perhaps millions—is to embed the experiment in a web service that's already being widely used, or package the experiment into a web-based advertisement. A growing number of researchers have made headlines with such covert experiments:

- In 2005, researchers at Indiana University sent targeted email messages to 921 members of the University community who became experimental subjects *who did not know that they were part of an experiment*. The email, which used email addresses that were spoofed from another 810 community members, directed the recipients to a website

[53] https://www.mturk.com/mturk/
[54] https://www.crowdflower.com/
[55] https://www.prolific.ac/
[56] The National Research Service Award Act of 1974, Public Law 93-348
[57] 45 CFR 46
[58] Marjorie Connley, "Push Polls, Defined," The New York Times, June 18, 2014. https://www.nytimes.com/2014/06/19/upshot/push-polls-defined.html

where their Indiana University username and password was requested. After the password was verified, the community members were told that they had been successfully "phished" and were directed to another website where they were provided with security training. Even though the study had been approved by the university's IRB and its computer security team, the timing of the experiment at the end of the semester and the fact that subjects were involved without their permission resulted in substantial negative publicity.[59]

- In 2010, Facebook conducted a study on an astounding 61 million of its users to see if it could influence the outcome of an election by selectively mobilizing different segments of its user population to vote. Facebook's researchers theorized that by showing clickable buttons in a user's newsfeed a clickable button saying "I Voted," a user's friends would be incentivized to vote because they knew that their trusted friends had voted. After it conducted the research, Facebook concluded that it could influence the outcome of an election by selectively showing the "I Voted" button or elements of the news feed to the specific subgroups.[60]

- In 2012, researchers at Facebook manipulated the "News Feed" feature of 689,003 Facebook users to see if such manipulations could alter the users' affective outlook. The researchers wanted to see if they could facilitate the transference of an emotional state from one user to another. In fact, Facebook could. There was significant negative publicity after the study was published.[61]

- Between May 2014 and January 2015, researchers at Georgia Tech and Princeton purchased online advertisements that delivered an experimental payload 141,626 times to 88,260 distinct IP addresses in 170 countries, resulting in more than 1,000 measurements each in China, India, the United Kingdom and Brazil, and more than 100 measurements each in Egypt, South Korea, Iran, Pakistan, Turkey and Saudi Arabia. The purpose of the experiments was to measure the pervasiveness of web censorship on a country-by-country basis, and this measurement was performed by instructing the unwitting users' web browsers to attempt to visit sensitive or controversial web content

---

[59] For a writeup of the experiment, see Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. *Commun. ACM* 50, 10 (October 2007), 94-100. DOI=http://dx.doi.org/10.1145/1290958.1290968. See also P. Finn and M. Jakobsson, "Designing ethical phishing experiments," in *IEEE Technology and Society Magazine*, vol. 26, no. 1, pp. 46-58, Spring 2007.

[60] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle & James H. Fowler, A 61-million-person experiment in social influence and political mobilization, *Nature*, Vol. 489, p. 295, September 13, 2012. http://fowler.ucsd.edu/massive_turnout.pdf

[61] The original study was published at Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788-8790. A comprehensive analysis of the event can be found in Jukka Jouhki, Epp Lauk, Maija Penttinen, Niina Sormanen and Turo Uskali, Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics, *Media and Communication*, Vol 4, No 4 (2016) http://dx.doi.org/10.17645/mac.v4i4.579

that was known to be blocked for moral or political reasons in some of these countries.[62] The research received considerable notoriety, as some of the researcher's colleagues believed that initiating such requests from web browsers in some countries might expose the users to undue risk from their governments. This was even more problematic because the users did not know that they were participating in a US-sponsored research experiment. This study was so controversial that the editors of the conference proceedings in which the article appeared felt compelled to include a boxed disclaimer statement on the first page of the article (see Fig. 1)

Statement from the SIGCOMM 2015 Program Committee: The SIGCOMM 2015 PC appreciated the technical contributions made in this paper, but found the paper controversial because some of the experiments the authors conducted raise ethical concerns. The controversy arose in large part because the networking research community does not yet have widely accepted guidelines or rules for the ethics of experiments that measure online censorship. In accordance with the published submission guidelines for SIGCOMM 2015, had the authors not engaged with their Institutional Review Boards (IRBs) or had their IRBs determined that their research was unethical, the PC would have rejected the paper without review. But the authors did engage with their IRBs, which did not flag the research as unethical. The PC hopes that discussion of the ethical concerns these experiments raise will advance the development of ethical guidelines in this area. It is the PC's view that future guidelines should include as a core principle that researchers should not engage in experiments that subject users to an appreciable risk of substantial harm absent informed consent. The PC endorses neither the use of the experimental techniques this paper describes nor the experiments the authors conducted.

Figure 1: Statement that appeared on the first page of "Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests," by Burnett and Feamster, 2015.

In 1979, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research published the *Belmont Report,*[63] named after the Belmont Conference Center in Elkridge, Maryland where the report was drafted. Written largely in reaction to the public disclosure of the Tuskegee Syphilis Study,[64] The *Belmont Report* laid the ethical groundwork for what became the National Research Act[65] and the Common Rule.[66] Today many

[62] Sam Burnett and Nick Feamster. 2015. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication* (SIGCOMM '15). ACM, New York, NY, USA, 653-667. DOI: http://dx.doi.org/10.1145/2785956.2787485
[63] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education and Welfare (DHEW) (30 September 1978). *The Belmont Report* (PDF). Washington, DC: United States Government Printing Office.
[64] "Tuskegee Study - Timeline". *NCHHSTP*. CDC. Page last reviewed December 22, 2015. Page last updated December 8, 2016. Page regtrieved December 17, 2016.
[65] The National Research Service Award Act of 1974, Public Law 93-348
[66] 45 CFR 46

federally funded-researchers in the United States who perform human subjects research are required under the Common Rule to trained and tested on the Report's three fundamental ethical principles: Respect for persons; Beneficence; and Justice.

In recognition that ICT research might require additional ethical principles, between 2010 and 2012 the US Department of Homeland Security convened a series of workshops with leading experts in the computing field to create a new report regarding the ethical conduct of research in the cyber age. Called *The Menlo Report*,[67] the new guide expanded the Belmont Report's original three ethical principles to include a fourth principle, "Respect for Law and Public Interest." This principle was added in recognition of the fact that experimentation on modern computer systems could result in significant damage to both those systems and the greater society.[68]

Key Points:

- The same technology trends that make it easy to work with organic data make it easy to experiment on large numbers of people without their permission.

- Such experiments can have significant impact on real-world events, such as influencing the outcome of elections and changing the emotional state of millions of people.

- Few mechanisms exist for regulating large-scale experiments outside of the Institutional Review Board (IRB) system, and there are many concerns regarding the ability of the IRB system to handle these kinds of experiments.

Conclusion

In the past, social scientists largely confined their work to datasets that they either made themselves or that they acquired from official statistics agencies. Today there is a growing interest in using organic data that results from administrative or operational systems. These datasets promise visibility into aspects of the economy and society that were never before available to social scientists, and offer resolution that was unimaginable until now. But these datasets also come with the risk of significant privacy violations and unknown data quality.

Techniques developed by computer scientists over the past four decades offer the promise of being able to work with these new datasets in a manner that is ethically appropriate and scientifically defensible. But these techniques, although they have been mathematically demonstrated and published in the peer reviewed literature, have not been developed to the point that they can be readily incorporated into production systems with manageable costs.

---

[67] D. Dittrich and E. Kenneally, "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research", Tech. rep., U.S. Department of Homeland Security, Aug 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/

[68] For example, in 1988 Robert Tappan Morris created a computer worm that was designed to assess the size of the Internet, but resulted instead in disabling more than 6,000 Internet-connected computers, roughly 10% of the computers that were connected to the Internet at the time. See Cyberpunk: Outlaws and Hackers on the Computer Frontier, by Katie Hafner and John Markoff, Simon and Schuster, Nov 1, 1995. 396 pages.

Social scientists need to work with computer scientists to move these techniques from the laboratory into practice. The benefits of doing so will be clear.